DOCUMENT RESUME

ED 415 627 EC 306 107

| | |
|---|---|
| AUTHOR | Langenfeld, Karen; Thurlow, Martha; Scott, Dorene |
| TITLE | High Stakes Testing for Students: Unanswered Questions and Implications for Students with Disabilities. Synthesis Report 26. |
| INSTITUTION | National Center on Educational Outcomes, Minneapolis, MN.; National Association of State Directors of Special Education, Alexandria, VA.; Council of Chief State School Officers, Washington, DC. |
| SPONS AGENCY | Special Education Programs (ED/OSERS), Washington, DC. |
| PUB DATE | 1997-01-00 |
| NOTE | 50p. |
| CONTRACT | H159C50004 |
| AVAILABLE FROM | National Center on Educational Outcomes, University of Minnesota, 350 Elliott Hall, 75 East River Road, Minneapolis, MN 55455; phone: 612-626-1530; fax: 612-624-0879; World Wide Web: http://www.coled.umn.edu/NCEO (document may be copied without charge, additional print copies $10). |
| PUB TYPE | Information Analyses (070) -- Reports - Descriptive (141) |
| EDRS PRICE | MF01/PC02 Plus Postage. |
| DESCRIPTORS | *Academic Standards; Accountability; Cost Effectiveness; Curriculum Development; *Disabilities; *Educational Assessment; Educational Environment; Educational Improvement; *Educational Testing; Elementary Secondary Education; Evaluation Methods; Learning; Outcomes of Education; *Student Evaluation; Student Participation; Teacher Attitudes; *Testing Problems |
| IDENTIFIERS | *High Stakes Tests |

ABSTRACT

This report reviews existing research on the effects of high stakes tests on students, particularly students with disabilities. The review focuses on potential effects on the curriculum, student learning, attitudes and school climate, and the costs versus benefits of high stakes testing of students with disabilities. Results indicate that research results on high stakes testing are inconclusive and vary with the type of research questions asked and the types of tests examined. The evidence suggests that teachers change the curriculum based on the tests and concentrate time and effort teaching to test content and format. The effects on student learning are largely unknown, but the evidence does suggest that increasing test scores in themselves do not serve as evidence that students are learning more. High stakes testing seems to have a negative effect on the attitudes and workloads of teachers, but little is known about the effects on students. States still do not take into account the full costs of high stakes testing programs, and claims that testing alone can cause major educational improvement have not been proven. Recommendations for future research are provided. An appendix summarizes the literature reviewed for the report. (Contains 46 references.) (CR)

Synthesis Report 26

ED 415 627

# High Stakes Testing for Students: Unanswered Questions and Implications for Students with Disabilities

EC 306107

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
This document has been reproduced as received from the person or organization originating it.
Minor changes have been made to improve reproduction quality.
Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

*In collaboration with:*

Council of Chief State School Officers (CCSSO)

National Association of State Directors of Special Education (NASDSE)



BEST COPY AVAILABLE

Synthesis Report 26

# High Stakes Testing for Students:
# Unanswered Questions and Implications for Students with Disabilities

Karen Langenfeld • Martha Thurlow • Dorene Scott

January 1997

NATIONAL CENTER ON EDUCATIONAL OUTCOMES

The Center is supported through a Cooperative Agreement (#H159C50004) with the Division of Innovation and Development, Office of Special Education Programs, U.S. Department of Education. Opinions expressed herein do not necessarily reflect those of the U.S. Department of Education or Offices within it.

**NCEO Core Staff**

Robert H. Bruininks
Judith L. Elliott
Ron Erickson
Dorene L. Scott
Patricia Seppanen
Martha L. Thurlow, Associate Director
James E. Ysseldyke, Director

Additional copies of this document may be ordered for $10.00 from:

National Center on Educational Outcomes
University of Minnesota • 350 Elliott Hall
75 East River Road • Minneapolis, MN 55455
Phone 612/624-8561 • Fax 612/624-0879
http://www.coled.umn.edu/NCEO

## Executive Summary

Increasingly, schools are administering tests that have important consequences for students. This literature review looks at existing research on the effects of these high stakes tests on students, with particular attention to students with disabilities. The review focuses on potential effects on the curriculum, students, student learning, attitudes, school climate, and costs.

Based on this review, we propose the following recommendations for future research and for those people who develop, implement, and evaluate large-scale testing programs:

- Focus on the effects of high stakes testing on students rather than on schools and systems.
- Assess the effects of high stakes testing on the curriculum for both special and regular education.
- Assess the effects of high stakes testing on the dropout rates for both special and regular education.
- Study the effects of high stakes testing programs on students who are excluded from testing.
- Develop assessments that are more inclusive of students with disabilities (for example, including students with disabilities in state norming samples, and norming tests with some students using common accommodations such as extended time and Braille).
- Study the effects of high stakes testing on the relationship between regular and special education.
- Develop a framework for evaluating the costs versus the benefits of high stakes testing programs, particularly for alternative and authentic assessments.

## Table of Contents


High Stakes Testing ..... 1
- What is High Stakes? ..... 1

Literature Review Procedure ..... 2

Curriculum and Instruction ..... 3
- What Effect Does High Stakes Testing Have on the Curriculum? ..... 3
- How are Educational Opportunities for Persons with Disabilities Affected by High Stakes Testing? ..... 5
- Directions ..... 6

Student Learning ..... 7
- How Well Does High Stakes Testing Measure Student Learning? ..... 7
- How Does High Stakes Testing Affect the Learning of Students with Disabilities? ..... 11
- How Does Inclusion of Students with Disabilities Affect State and Local Test Scores? ..... 12
- How Do Students with Disabilities Perform on High Stakes Assessments? ..... 13
- Directions ..... 15

Attitudes and School Climate ..... 17
- What Effect Does High Stakes Testing Have on Teachers' and Students' Attitudes, and on the Climate of Learning? ..... 17
- What are the Emotional/Attitudinal Effects on Students with Disabilities? ..... 18
- Directions ..... 19

Costs Versus Benefits ..... 19
- What are the Costs vs. the Benefits of High Stakes Testing? ..... 19
- Who Benefits the Most from High Stakes Testing? ..... 21
- What About Portfolio and Authentic Assessments? ..... 22
- What are the Costs vs. the Benefits of High Stakes Testing for Students with Disabilities? ..... 25
- Directions ..... 25

Recommendations and Conclusions ..... 25

References ..... 29

Appendix: Summary of Literature Reviewed for This Report ..... 35

## High Stakes Testing

Over the past two decades, statewide testing of students has become more and more common. This testing is conducted to meet a multitude of purposes and a wide variety of instruments is used. The practice of statewide assessment has been the center of a great deal of debate and political rhetoric (Popham 1987; Salganik, 1985; Shepard 1992). Although some states have attached high stakes, such as graduation, to tests since the early 1970s (e.g., Florida, New York), research findings on the effects of high stakes testing have not been decisive. The Office of Technology Assessment (OTA) (1992) summarized the findings by stating:

> In the end, then, there appears to be consensus that innovation in school testing policies can have profound effects–the disagreement is over the desirability of those effects. Although some of the evidence is contradictory, at times even confusing, one thing is clear: test-based accountability is no panacea. Specific proposals for tests intended to catalyze school improvement must be scrutinized on their individual merits (p. 15).

Research looking at effects on students, particularly students with disabilities, is the focus of this report. In the past, most research has emphasized school improvement and system level effects. It is important, therefore, to carefully examine the research for evidence of student level effects.

As the clients of our educational system became more diverse in their characteristics (Hodgkinson, 1992; Hodgkinson & Outtz, 1992), it becomes increasingly important to look more specifically at differential effects of major reforms. Since we know that groups of students perform differently (Mullis et al., 1994), testing, particularly high stakes testing that may have important consequences for students, has the potential to affect some groups differently from other groups. Among the groups of students for whom differential effects might be expected are those with disabilities (Olson & Goldstein, 1996). At present there is increased pressure for states to use tests to decide whether students will earn a high school diploma and to hold all students to the same requirements for graduation (see Thurlow, Ysseldyke, & Anderson, 1995). It is important to re-examine what we know about high stakes testing and what the research literature tells us.

### What is "High Stakes"?

A test can be considered high stakes if the results of the test have perceived or real consequences for students, staff, or schools (Madaus, 1988). Increasingly, states, cities, and school boards are using test scores in order to evaluate schools and allocate resources. In October 1996, Chicago put 109 schools on academic probation. According to Hendrie (1996)

scores from nationally normed standardized tests were a chief factor in determining who would be placed on probation. Manzo (1996) reported that Philadelphia was planning to link teacher raises and cash awards to schools based on student test scores, attendance, and graduation rates. For schools with chronically low-performing students, schools could be forced to replace up to three-fourths of their staffs.

The consequences of testing can be both intended and unintended. Corbett and Wilson (1991) state:

> Stakes can become high when test results automatically trigger important consequences for students or the school system, and also when educators, students, or the public perceive that significant consequences accompany test results. Thus, a formal trigger of consequences need not be built into the testing program for stakes to be high. Instead, test results can cause the public to make an assessment of the quality of the school system that serves them, and this judgment in turn can lead to a conclusion that children's choices . . . have been affected. The product of this process can be increased public pressure to improve test scores, especially when the perception is that the system is likely to have a negative impact on those choices. (p. 27)

While this review includes research using this broad definition of high stakes, the high stakes tests in which we are most interested are those that determine a student's progress through and out of school. Examples are: minimum competency tests (MCTs), graduation exit exams, and tests used to decide promotion from grade to grade.

## Literature Review Procedure

This analysis of the literature is based on educational and psychological journals and unpublished research from 1980 to the present. The literature search revealed fewer than thirty research studies, with only five studies specifically focusing on persons with disabilities. The articles include published and unpublished research, state-sponsored evaluation and research reports, papers from professional meetings, and articles from various research organizations such as CRESST (Center for Research on Evaluation, Standards and Student Testing).

It became clear, as we reviewed the available evidence, that we would generate more questions than answers. The review was limited not only by the lack of research in this area, but also by the quite broad uses of testing for high stakes purposes. For example, some of the studies focused on elementary schools and the pressure placed on schools to do well for financial or other purposes. Other studies focused on high school exit exams and still others focused on exams that determine promotion from grade to grade.

Since the articles contained information on a variety of situations that are considered "high stakes" (and thus are not directly comparable), a meta-analysis of the data was not attempted. Instead, a decision was made to conduct a descriptive analysis, and to organize the data as a series of research questions, providing summary and analysis of results and directions for future research. Most of the studies included here are based on state tests, yet each state has a different approach to testing, and each program is evolving. **The studies and evaluation reports used in this review do not necessarily reflect the particular state's current testing practices.** A high stakes testing program in which schools are evaluated according to the results on the Iowa Test of Basic skills is different from a program in which students must pass a test in order to graduate from high school. While often the studies themselves are not directly comparable, we do have enough studies in this review to see emerging patterns that require further investigation. A summary of the studies we used in this review is provided in the appendix.

This report is organized around a series of questions. The available literature addresses some of these questions, but all are in need of further exploration and consideration by the people who make testing decisions. The questions center around four basic effects of high stakes testing: effects on the curriculum, effects on student learning, effects on student and teacher attitudes and the climate of learning, and the costs versus the benefits of high stakes testing. Each section of this review focuses on high stakes testing in general, and then explores the implications for persons with disabilities. Directions for future research also are proposed.

## Curriculum and Instruction

Most research on the effects of high stakes testing has looked at the curriculum and instruction. Still, the information from this research is relatively indirect. Seven studies were reviewed (see first table in the appendix) that addressed the effects of high stakes testing on the curriculum and instruction.

### What Effect Does High-Stakes Testing Have on the Curriculum?

Several claims have been made as to the possible positive and negative effects of high stakes testing on the curriculum. Some individuals, such as Popham (1987), believe that the curriculum will improve as schools, teachers, and students attempt to meet the challenges that testing will impose. Others (Madaus, 1988; Shepard, 1992) fear that high stakes testing will narrow the curriculum, focus on lower-order skills, or take control of the curriculum away from local sources.

In a study comparing a low-stakes state (Pennsylvania) and a high stakes state (Maryland), Corbett and Wilson (1990, 1991) found that in a high stakes situation, teachers reported a narrowing of the curriculum, but that not all teachers thought this was a bad thing. According to Corbett and Wilson (1990), "Maryland school districts focused more directly on improving test scores, altered the curriculum to a greater extent, reported more improvement in the curriculum, and felt the curriculum had narrowed more than their Pennsylvania colleagues" (p. 72). They also found that smaller districts were more likely to make greater curriculum and instruction adjustments. Teachers did not always think that these changes were necessarily in the best interests of the students. In general, high stakes testing affected both the content and the sequence of instruction, and efforts to affect test scores directly increased as the testing dates approached.

Rottenberg and Smith (1990) used qualitative interviews in a high stakes elementary setting in a state (not identified) that used the Iowa Test of Basic Skills (ITBS). The test was considered high stakes because the scores were used in the evaluation of principals and for making curriculum decisions. The media also reported scores on the ITBS by school and grade level. Rottenberg and Smith found that testing reduced the time available for ordinary instruction. Schools were also neglecting material not in the tests, while encouraging the use of instructional methods resembling testing, such as multiple-choice exams.

Shepard and Dougherty (1991) addressed test-preparation practices and effects of testing on instruction. They used districts from two states with high stakes testing, one in the Southwest and one in the Southeast, and sampled teachers in the 3rd, 5th, and 6th grades. They found that teachers gave greater emphasis to basic skills instruction, and that nontested content suffered because of the focus on standardized tests. They also found that teachers spent an inordinate amount of time preparing for tests rather than focusing on the regular curriculum. They reported four weeks of intensive test preparation, plus two weeks for administering the test itself. This emphasis on preparation was not limited to the time surrounding the test administration: 68% of the teachers reported using worksheets throughout the year to review expected test content and to give students practice with the testing formats. Consistent with other studies, teachers did report that the tests helped them to set clear instructional goals.

Herman and Golan (undated) also found that teachers spent an inordinate amount of time preparing for tests. They surveyed upper elementary school teachers in nine states with high stakes testing using a 136 item questionnaire designed specifically for this study. Consistent with Shepard and Dougherty (1991), they found that teachers were spending class time on worksheets covering test content and format. Teachers also changed the content and sequence of instruction throughout the year. While higher order skills and nontested subjects such as arts and science had less emphasis, teachers continued to cover these subjects. Basic skills,

however, were given the most emphasis. As a result of testing, teachers changed the content and sequence of instruction based on prior tests and how their classes did the previous year. Another trend found in this study was that socioeconomic status (SES) had a great deal to do with how well students performed on the tests and how much the tests affected instruction. The authors found that socioeconomic status was significantly and negatively related to the amount of attention that schools and teachers gave to test scores, curriculum planning, and time devoted to test-related activities. According to Herman and Golan (undated), "testing is more influential and exerts stronger effects on teaching in schools serving more disadvantaged students" (p. 58).

Rodgers, Paredes, and Mangino (1991) looked at the effects of the Texas Educational Assessment of Minimum Skills (TEAMS), a test that students needed to pass in order to graduate from high school. The study took place over five years, using 12,404 eleventh grade students from the Austin Independent School District. The test focused on language arts and math. Rodgers et al. found that basic skills, as measured on the Tests of Achievement and Proficiency (TAP), increased as a result of the minimum competency exam, but that higher order skills remained the same. They concluded that districts should be cautious about narrowing the curriculum and letting higher order skills suffer for the sake of improving test scores.

In summary, the literature provides very little direct evidence about the effects of high stakes testing on the curriculum. The results are consistent, however, in showing that high stakes testing does affect what and how teachers teach. One can argue about whether these effects are positive or negative. For example, Shepard and Dougherty (1991) found that high stakes testing narrowed the curriculum; however, in Corbett and Wilson (1990), some teachers thought that the curriculum had narrowed while others thought that the curriculum had become more focused. Berger and Elson (1996) found that in a high stakes environment, teachers reported a clearer mission for their schools. The authors claimed that this supports advocates' claims that measurement driven instruction "adds focus, coherence, and clarity to the mission of a school. Teachers and students know what is expected of them and how they will be judged" (p. 22). These may all be aspects of the same phenomenon and deserve to be looked at more closely.

### How Are Educational Opportunities For Persons With Disabilities Affected By High Stakes Testing?

We could find no studies that directly assessed the effects of high stakes testing on the curriculum for students with disabilities. Indeed, some researchers, such as Rodgers et al. (1991), purposely excluded students with disabilities from their studies in order to make data

analysis easier. This section addresses some of the evidence suggesting that educational opportunities for students with disabilities are negatively affected by high stakes testing.

Bergquist, Elzie and Groves (undated) conducted an evaluation of the impact and effectiveness of Florida graduation and competency test standards on students with disabilities in 1986-87. They found that students with disabilities had trouble earning the 24 credits required for a standard diploma in four years. This made it difficult for them to accommodate vocational training in their programs. Students with disabilities even had trouble earning 24 credits toward a special diploma with courses paralleling the standard diploma, and these students also had difficulty including vocational training in their programs. The authors concluded that students were "more likely to leave high school with a standard or special diploma but no marketable job skill" (p. 10). Also, accommodations for students in the academically-oriented classes varied from district to district. As a result, children who failed in one district might have succeeded in another district that was more willing to meet their needs. The authors reported that "districts with greater flexibility in the course requirements for the diploma were better able to meet the unique needs of the handicapped student through the course offerings at the high school level" (p. 10).

Little is known about the effects of increased graduation standards on curriculum offerings. Grossman, Kirst, and Schmidt-Posner (1986) found that increased graduation requirements and entrance requirements for college in California resulted in increased offerings in math and science, and decreased offerings in industrial arts, home economics, business education, etc. If this is a national trend, then one can expect that as resources decrease and become more concentrated on academic subjects, the opportunities for students to find an appropriate education may decrease as well.

## Directions

It is interesting that the majority of the research is almost entirely focused on teachers and their perceptions rather than on students and their performance. For example, do students study longer for tests when the stakes are high? If they do, does this generalize to the rest of the curriculum, or are students more likely to study only for the test, and what they believe the test will include?

For students with disabilities, it is especially important to assess how high stakes testing affects the curriculum. Should students be concentrating on vocational and other objectives, or should their IEPs focus on the skills measured on a high stakes test? Educators and legislators must be made aware that the needs of students with disabilities may not always be best met with what is offered in the regular academic curriculum. If students with disabilities are to be

included in educational reforms, then their needs both for accommodations in testing and modification of the curriculum must also be taken into account.

The evidence reviewed so far suggests that high stakes testing affects the curriculum. These effects can be viewed as positive or negative, but the negative view is more pervasive. It is equally, if not more important, however, to ask what effect these tests have on students and learning.

## Student Learning

Indeed, questions about the impact of high stakes testing on students with disabilities are the critical ones that need answering. To date, however, there is virtually no evidence that adequately addresses the question of how high stakes testing affects student learning. Again, we turn to the study by Corbett and Wilson (1991) of a high stakes state (Maryland) where teachers spent more time on test preparation, used more practice tests, and conducted more content reviews. They noted, however, that "while the numbers visually document more intense activity on the part of educators in Maryland, missing is any assessment regarding its value to improved learning" (p. 70).

The media have touted many statewide testing programs on the basis that students' scores on the tests have improved over time. In this section, we review studies (see second table in the appendix) that suggest this interpretation of test scores as a measure of actual student improvement may be premature. Many factors that have nothing to do with actual learning can affect scores on high stakes tests. This section addresses several of these factors, including teaching to the test, reclassifying students so that they no longer have to take the test (e.g., through referral to special education), and increased dropout rates.

### How Well Does High Stakes Testing Measure Student Learning?

"Teaching to the test" means a concentration on skills that increase test scores *regardless* of the amount of knowledge the student actually possesses. There are some generally accepted skills, often called "testwiseness," that many states openly encourage. Testwiseness includes things such as getting a good night's sleep, knowing how to make educated guesses, using relaxation techniques, etc. Other practices are considered unethical and include such things as giving students copies of the test prior to administration, hinting at answers, and changing answers on test papers. In the middle ground, there exist practices such as giving students worksheets (multiple choice, five-minute essays) to familiarize students with the format of the tests. One extreme example of this was reported by Madaus (1988) of an entire course dedicated to five-

paragraph timed argumentative essays. This course was being offered in lieu of other advanced courses in writing.

Several studies have been conducted to determine whether increases in test scores reflect real increases in student learning. Shepard (1990) conducted interviews with teachers in a high stakes testing environment to investigate the claim that "teaching to the test" was causing inflated test scores. Most instances of teaching to the test were of a mild form, such as using commercially marketed programs designed to teach test-taking skills to students. Other practices included such things as school-developed practice tests, pep rallies to "psych kids up" to do well, and high school courses specifically aimed at the competency measures. This study found evidence that questionable practices existed, but could not quantify their effects on actual test scores.

In a different approach, Walstad (1984) looked at what kinds of practices were responsible for increases in test scores. Controlling for other factors such as socioeconomic status (SES), Walstad looked at three variables: pretesting students, curriculum changes based on the state's education standards, and district-sponsored workshops to increase the skills of teachers in implementing the standards. Pretesting, a practice where students were able to practice the test format, was the *only* significant variable that contributed to an increase in test scores. Curriculum and instructional changes had no significant impact. This suggests that increases in the test scores were not due to actual learning, but rather to familiarity with the tests.

Another way to determine whether high stakes tests are measuring actual learning is to look at how they generalize to other tests. In theory, if a student is learning a skill such as spelling, this knowledge should generalize across state tests and other tests of achievement. Koretz, Linn, Dunbar, and Shepard (1991) looked at whether performance on a third grade high stakes test would generalize to other tests. Mathematics scores did not generalize from one test to another, and reading scores generalized only a small amount, of little practical significance. Koretz et al. concluded that "to a substantial degree, teachers in this district must be focusing on content that is specific to the particular test used for accountability, rather than trying to improve achievement in the broader sense that we would all desire" (pp. 20-21).

Other factors that have nothing to do with learning also may influence test scores. Shepard (1992) stated that:

> Because of the pressure on test scores, more hard-to-teach children are rejected by the system. There is a direct correspondence between accountability pressure and the number of children denied kindergarten entrance, assigned to two-year kindergarten programs, referred to special education, made to repeat a grade, or who drop out of school (p. 5).

This kind of academic "redshirting"—keeping students back a grade in order to improve test scores (see Zlatos, 1994) has been found in several other studies as well. Potter and Wall (1992) found evidence that, as early as preschool, children were kept back a grade so that they would do better on the tests. Allington and McGill-Franzen (1992a) examined test scores in districts that had claimed increases in student performance on high stakes tests. The districts came from a variety of settings (urban, suburban, rural) and socioeconomic status. Rather than finding evidence of increased learning and better teaching, the authors found an increase in the proportion of students retained a grade or placed in special education. The authors recalculated the test data by determining which children started kindergarten together. When test scores of children who had been identified for special education or who had been held back a year were included in the test scores, the gains districts had been reporting disappeared.

Allington and McGill-Franzen (1992b) also looked at trends in these schools and found an increasing number of students were being referred to special education or retained a grade during a period of increased high stakes.

Similar factors also were found to influence test scores on the Texas minimum competency tests of the mid-eighties. Mangino, Battaile, and Washington (1986) found that increased test scores were probably due to factors other than actual learning. They identified several problem areas, including students taking the tests many times, the ability of students to gain waivers from taking the tests, and a higher percentage of students using special education exemptions. Thus, a school with incentives to do better on the statewide tests could improve overall scores by encouraging poor students to get waivers, or by placing them in special education.

Dropping out of school is perhaps the most serious of the effects of high stakes testing, but it has been very difficult to prove a causal connection between high stakes testing and dropout rates. Indirect evidence of increased dropout rates has been found by researchers such as Potter and Wall (1992) who looked at the effects of a graduation exit exam on students in South Carolina. They found that more students were retained a grade as a result of the high stakes testing. Overage students were more likely to drop out of school. In a longitudinal study from 1982-1989, Morris (1991) looked at patterns of changes in grade retention rates in a large urban school district in Florida, where tests were considered high stakes due to pressure from the media. Among other factors (such as restructuring the schools from a junior high to a middle school model), he found that high stakes testing and increased graduation standards increased the retention rates. He found that retention rates increased during the test years when the state introduced a diagnostic test to identify failing students for remediation. The increases in retention rates spread to other grades when the test was used to identify weak school programs. When the graduation requirements increased while the time to achieve the requirements decreased, retention rates also rose.

Another study that attempted to establish a link between high school dropout rates and high stakes testing was conducted by Catterall (1987). This study looked at states that had minimum competency examinations for graduation from high school. Catterall used interviews to ask teachers and students their perceptions of the impact of minimum competency testing on dropout rates. Few teachers and administrators thought that minimum competency testing had much of an impact, and none could cite any evidence to support their beliefs. They saw the minimum competency tests as "largely meaningless and innocuous" (p. 22). Students, however, had very different perceptions, and 14% of the students said that they knew someone who had failed the test and dropped out of school as a result. Catterall also found a high correlation between students who had failed the test at least once and who expressed doubt as to whether they would finish school.

Two recent studies have attempted to examine the link between high stakes testing and dropout rates directly, with very different results. The studies looked at different populations and types of tests. Griffin and Heidorn (1996) reported on the Florida minimum competency tests in mathematics and communication, which students must pass in order to graduate from high school. Students may take the tests up to five times, starting in grade 10. The authors found that dropout rates increased only for students who were doing well academically and subsequently failed the tests. Dropout rates did not increase for students who already had poor academic records, or for minority students.

Reardon (1996) found very different results when he examined minimum competency tests that students needed to pass in order to be promoted from eighth to ninth grade. This study used data from the National Educational Longitudinal Study (NELS) to look specifically at retention and dropout rates related to high stakes testing. Interestingly, Reardon found that minimum competency testing was more prevalent in urban schools with high concentrations of low-income and minority students.

> This uneven distribution of MCT requirements may simply mean that the prevalence of MCTs is related to the prevalence of lower-achieving students–the group proponents believe the tests are most likely to help. But it raises an important concern as well: if MCTs do influence some students to drop out who would not have otherwise, then not only are MCT policies harmful, but their harmful effects are disproportionately concentrated on those students with the fewest opportunities for success (p. 5).

Reardon did, in fact, find evidence that dropout rates do increase as a result of minimum competency testing and that furthermore, "the . . . data also suggests that it is the concentrated poverty of these schools and their communities, and their concomitant lack of resources, that link MCT policies to higher dropout rates, rather than other risk factors, such as student grades, age, attendance, and minority group membership" (p. 5). Reardon found that eighth

grade students with minimum competency testing requirements dropped out at double the rate of students without MCT requirements (8.8% as opposed to 4.2%). Breaking down the data by socioeconomic status, Reardon found that low and moderately low SES schools were the most related to high dropout rates. MCTs had little or no effect on the dropout rates in higher SES schools.

While these results conflict with those of Griffin and Heidorn (1996), these differences can be explained by the differences in the type of tests and populations used in the studies. The Florida test, on which Griffin and Heidorn base their study, has been described as a basic-level test that does not set particularly high standards. In addition, students may take the test up to five times, perhaps relieving some of the pressure of the high stakes. The Reardon (1996) study, on the other hand, uses a more representative sample, concentrating on high stakes tests that have earlier and more immediate consequences for students.

In summary, increased test scores on a high stakes test do not necessarily translate into increased learning for students. The problems go beyond simple inflation of test scores, which is a serious concern from a measurement standpoint. These studies point to a frightening but very real possibility that children will be systematically and deliberately labeled, excluded, and pushed out of the system altogether in order to improve test scores. Madaus (1988) argued that teaching to the test is part of human nature.

> Some have argued strongly that if the skills are well chosen, and if the tests truly measure them, then coaching is perfectly acceptable. This argument sounds reasonable, and in the short term, it may even work. However, it ignores a fundamental fact of life: when the teacher's professional worth is estimated in terms of exam success, teachers will corrupt the skills measured by reducing them to the level of strategies in which the examinee is drilled . . . . The view that we can coach for the skills apart from the tradition of test questions, embodies a staggeringly optimistic view of human nature that ignores the powerful pull of self-interest. (p. 93)

## How Does High Stakes Testing Affect the Learning of Students with Disabilities?

Macmillan, Balow, Widaman, and Hemsley (as cited in Griffin and Heidorn, 1996) found that students with disabilities who failed a minimum competency exam had higher dropout rates than regular education students. This study, however, did not control for other factors that might also affect dropout rates, such as academic performance or problem behavior.

As far as effects on what students with disabilities are actually learning as a result of high stakes testing, we do not really know. If researchers have paid little attention to effects on typical students, they have paid even less attention to students with disabilities. Even where

students with disabilities are included in statewide testing programs, they are often left out of research and evaluation reports. This is unfortunate, since people with disabilities have much to lose when it comes to high stakes testing, especially for graduation purposes. As Safer (1980) pointed out, students who lack a regular diploma may be discriminated against in employment later in life.

## How Does Inclusion of Students with Disabilities Affect State and Local Test Scores?

Inclusion of students with disabilities in statewide assessments causes problems for many persons concerned with the accurate measurement of students. When persons with disabilities are not included in norming samples, or accommodations such as large print editions are not part of the initial test standardization process, then including people, or allowing accommodations raises legitimate questions about the reliability and validity of the test scores. According to Bond, Roeber, and Braskamp (1996), 31 states currently use norm-referenced tests, such as the Iowa Test of Basic Skills (ITBS) or the Stanford Achievement Test (SAT). Scores on these tests are used for improvement of instruction (29 states), program evaluation (24 states), school performance reporting (22 states), student diagnosis (19 states), and high school graduation (two states). Many times, students with disabilities are left out of these tests or denied accommodations because the test was not standardized on these populations. Current research efforts are trying to determine whether accommodations such as large print, extra time on tests, Braille versions of tests, etc. are a valid means of including students in these assessments. Other efforts have been made to determine whether tests can be administered to persons with disabilities and reported using separate norms and percentile ranks. Chin-Chance, Gronna, and Jenkins (1996) found that the state of Hawaii had enough students taking the state's norm-referenced test to report separate norms and percentile ranks for students with and without disabilities and within disability categories.

Research into testing for students with special needs should not stop at simply determining the validity of using a type of accommodation. It is possible to develop tests that are normed on all students, and using formats that would include a greater number of people. For example, tests can be developed using larger print, so that a separate test in large print need not be developed. Test items can be developed keeping in mind that some people may not be able to see the figures, but will need to rely on verbal information.

Criterion-referenced tests, for which a student must simply demonstrate a proficiency in a subject are more easily adaptable for persons with disabilities. Here, the student need only demonstrate proficiency in the areas to be tested. For example, in the state of Hawaii, students must pass the Hawaii State Test of Essential Competencies (HSTEC) to receive a high school diploma. Students who fail the HSTEC once may take the test again until they pass, or may

pass the test through the Essential Competencies Certification Center that uses an open-ended response format. According to Bond et al. (1996), 36 states currently use criterion-referenced tests. Sixteen states use criterion-referenced tests for graduation purposes. Thirty-four states use written assessments, 12 of them for graduation, and 18 states use alternative forms of assessment, two of them for graduation. While inclusion of students with disabilities in these types of assessments is more easily accomplished than in norm-referenced tests, it is still imperative that states take the needs of students with disabilities into account when they develop testing programs in order to avoid costly problems later.

## How Do Students with Disabilities Perform on High Stakes Assessments?

In general, the few studies that have been conducted on students with disabilities show that they do poorly when compared to peers without disabilities. In Hawaii, Chin-Chance et al. (1996) looked at the scores of students with disabilities taking the Stanford Achievement Test 8th Edition (SAT8) in the 3rd, 6th, 8th, and 10th grades. Hawaii uses the SAT8 for national, school, district, and local comparisons. While the study was not conducted using the state's graduation exit exam (the Hawaii State Test of Essential Competencies), the SAT8 can still be considered high stakes because the scores are public, comparisons are made among districts and schools, and SAT8 scores are used to make judgments about a school's effectiveness. The study looked at data for students with disabilities who took the statewide administration of the SAT8 in 1994 and 1995. They found that students with disabilities in all categories did more poorly on the tests than did students without disabilities; however, when looking at difference scores from one year to the next, students in Hawaii showed more improvement than did the national norms, and students with disabilities did as well as, or better than, students without disabilities. This study did not look at how students with disabilities were faring on the graduation exit exam, a test with much higher stakes than the SAT8.

Safer (1980) found that in Florida in the late 1970s, students with disabilities were not likely to pass the minimum competency test required for graduation (see Table 1).

According to the data reported by Safer, between six percent (educable mentally retarded) and 71% (speech/language impaired) of students passed the communications subtest. Altogether, only 46% of the special education students taking the Florida graduation exit exam passed the test. Between one percent (educable mentally retarded) and 33% (speech/language impaired) passed the math subtest, only 18% of the total.

| Table 1. Percentage of High School Juniors with Disabilities Passing Two Subtests of the Florida Minimal Competency Examination in 1977. | | | |
|---|---|---|---|
| **Handicapping Condition** | **Communications Subtest** | **Math Subtest** | **N** |
| Speech/language impaired | 71 | 33 | 509 |
| Deaf | 47 | 18 | 126 |
| Hard of hearing | 65 | 29 | 49 |
| Physically impaired | 67 | 14 | 110 |
| Emotionally disturbed | 56 | 17 | 114 |
| Socially maladjusted | 49 | 25 | 79 |
| Learning disabled | 49 | 17 | 502 |
| Educable mentally retarded | 6 | 1 | 479 |

From Safer (1980), p. 289

McKinney (1983) looked at how students with disabilities performed on the North Carolina Minimum Competency test, both with and without accommodations. He found that some groups were more likely to benefit from test modifications than others. Still, the probability of passing the test was low, especially for students with mild mental retardation, even with modifications. Out of 3,043 students taking the MCT in 1978, the reported pass rates (see Table 2) were only slightly better than those reported by Safer (1980). Even so, many students with disabilities could not pass both subtests, a requirement for receiving a diploma. Of the students who failed the test on the first try, 78% took the test a second time. Of these students, only 35% passed the reading subtest and 28% passed the math subtest. Students with visual impairments had the best retest success rates (72%) and students with mild mental handicaps had the lowest retest success rates (21%).

| Table 2. Percent of Students Who Passed the Subtests on the North Carolina Minimum Competency Test for Graduation. | | |
|---|---|---|
| **Handicapping Condition** | **Reading Subtest** | **Math Subtest** |
| Educable mentally handicapped | 12 | 7 |
| Learning disabled | 56 | 47 |
| Visually impaired | 92 | 88 |
| Hearing impaired | 75 | 70 |
| Multiply handicapped | 32 | 28 |
| "Other" handicapping conditions | 66 | 57 |

From information in McKinney (1983), p. 547-548

McKinney also looked at how test modifications affected student performance on the test. Half of the students in the sample received modified tests. Modifications included extended time, small group administration, audio cassettes, large print editions, and sign language. Use of modifications was not assigned at random, but varied across educational districts and type of school personnel (special education teachers were seven to 17 times more likely to use test modifications). Students with mild mental retardation who used modifications were more likely to pass the test than those who did not receive modifications. Students with hearing impairments were actually less likely to pass with modifications. Test modifications had no significant impact on other groups.

In interviews with school personnel, McKinney (1983) found that special educators were concerned about the impact of the testing on students with mild mental impairments. Teachers reported that "some exceptional students, particularly educable mentally handicapped students found the test extremely frustrating" (p. 549). Even though students with mild mental disabilities benefited from the accommodations, they were still less likely than any of the other groups to pass the tests, and were reportedly more likely to become frustrated by the testing process. Hall and Gallagher (as cited in Vitello, Camilli, and Molenaar, 1987) analyzed the same data and found that while remediation did make it possible for 50% of the students with disabilities retaking the test to pass on the second try, students with mild mental retardation were not helped by remediation efforts.

Vitello, Camilli, and Molenaar (1987) also found that students with mild mental handicaps had the most difficulty passing minimum competency exams. Their study examined the scores of the 4,299 students with disabilities who took the New Jersey competency test. At this time, students with disabilities who completed their IEP objectives were given a standard diploma. Of the students with disabilities who were eligible, only 40% actually took the exam, and 12% of these students passed. Only one percent (26 of 1438) students with mild mental disabilities took the exam, and, of this group, only four percent actually passed the test.

### Directions

The studies reviewed so far focus primarily on factors that may inadvertently influence test outcomes. The literature says very little about exactly what students should be learning or doing differently as a direct result of testing. More studies, such as that of Catteral (1987), which focus specifically on students and how testing reforms affect their lives and learning are also needed. The conflicting results of Griffin and Heidorn (1996) and Reardon (1996) indicate that not all testing reforms have the same results. It seems to us that every major testing reform requires the type of analysis that Griffin and Heidorn conducted, as well as continuing large-scale studies such as Reardon's which look at a broader spectrum of tests.

There is a dearth of evidence about the effects of high stakes testing on students with disabilities. If, for example, schools are referring more students to special education as a result of high stakes testing, then we need to know whether these students are getting the instruction they need, or whether we are simply putting them in positions where less is expected of them. The effects of both inclusion and exclusion of students with disabilities should be examined. For example, if we exclude students with disabilities, we may then spend less time teaching them important skills that the tests cover. If we include students with disabilities, we may then be neglecting other parts of their education, such as vocational skills that the IEP team believes to be important.

The exclusion of any group from a high stakes testing program makes it difficult to evaluate the effects on these students, particularly when looking at effects on dropout rates. When some students are excluded from testing in the first place, an increase in the dropout rate of this population would not show up in the results of studies such as Griffin and Heidorn (1996) or Reardon (1996). Students with disabilities, particularly learning and behavioral disabilities are already at a particularly high risk for dropping out of school (Sinclair, Christenson, Thurlow, & Evelo, 1994). It is possible that exclusion from high stakes testing–especially when the result is denial of a diploma–could push even more of these students out of school.

The evidence so far suggests that students with disabilities do not fare well on minimum competency tests. Furthermore, we still do not know whether preparing for the tests, taking the tests, or passing the tests have any consequences, positive or negative, for these students.

It is important for special educators to encourage new testing programs, and large testing companies, to begin including students with disabilities in norming samples, and to include some common accommodations (such as extended time, reading directions, large print editions) when developing tests. This will allow educators to more accurately measure the performance of students with disabilities, and increase the inclusion of students with disabilities in important assessment activities.

Thus far, we have examined the influence of high stakes testing in two obvious areas–curriculum and instruction, and student learning and performance. In those sections we found evidence that teachers focus the curriculum on the content and structure of tests. We also found that test scores often are corrupted through artificial inflation due to teaching to the test and reclassifying or retaining students who perform poorly so that they do not have to take the tests. High stakes testing may have other effects as well, effects that have nothing to do with academic performance or test scores.

## Attitudes and School Climate

This section addresses the less tangible issues of school climate and the attitudes of teachers and students toward testing (see the third table in the appendix for a summary of the literature reviewed). In this section, we distinguish between two types of high stakes testing: high stakes for students, and high stakes for schools and teachers. In most instances, if a test is high stakes for one group, it will be high stakes for all groups. Unfortunately, most of the research in this area has concentrated on the effects of high stakes testing on teachers and schools. High stakes for students include failure to get a diploma or grade promotion. Students may also experience high stakes when pressure is put on schools to perform better, and the schools in turn put pressure on the students. Schools and teachers may experience high stakes in terms of funding (for example, bonuses for good performance, decreases in funding for poor performance), public scrutiny by the media and politicians (for example, publishing test scores comparing schools or districts), or job security (for example, threatening to restructure schools if test performance does not improve).

### What Effects Does High Stakes Testing Have on Teachers' and Students' Attitudes, and on the Climate of Learning?

Corbett and Wilson (1991) found that staff in a high stakes state (Maryland) reported greater impact on their students' and their own lives than did staff in a low-stakes state. Teachers in the high stakes state also reported more stress, more paperwork, and decreased reliance on their professional judgment.

A qualitative study using classroom observations and interviews of teachers by Rottenberg and Smith (1990) found negative effects for both students and teachers in a high stakes testing program. They looked at the role of external testing in elementary schools in Arizona. The tests used in these schools, such as the Iowa Test of Basic Skills (ITBS), were considered high stakes because the results were used in the evaluation of principals and schools, and because the media reported ITBS scores by school and grade level.

> For pupils, particularly younger ones, most teachers believe that standardized testing is cruel and unusual punishment. Because of the length and difficulty of tests, the number of tests, the time limits, the fine print, and the difficulty in transferring answers to answer sheets, teachers believe tests cause stress, frustration, burn-out, fatigue, physical illness, misbehavior and fighting, and psychological distress. Some teachers believe that the tests cause their pupils to develop test anxiety and a failure mentality. (p. 17)

Effects on teachers were perceived to be equally negative:

> [Teachers] feel ashamed and embarrassed if their pupils score low or fail to grow by district standards. They feel relieved rather than proud when scores

> are high, for they know that test scores are weighted more by pupils' socioeconomic status and level of effort than anything teachers personally do in the classroom. (pp. 17-18)

Similar effects were found by Herman and Golan (undated) who reported that pressure to improve test scores was negatively related to job satisfaction and pride in teaching for upper elementary school teachers in nine states across the country.

A recent study by Berger and Elson (1996) looked at the effects of minimum competency testing on teacher autonomy, cooperation and school mission. They used data from the 1987 Schools and Staffing Survey conducted by the U.S. Department of Education that included surveys of over 19,000 teachers across the country. They compared teachers' responses in high stakes testing programs to those in states with low stakes testing programs. A high stakes program was defined as one in which the diploma is withheld when the student does not pass. As mentioned before, they found that high stakes were correlated with a clearer sense of mission for their schools. This generally positive finding, however, was accompanied by a reduced sense of autonomy. Contrary to other studies, Berger and Elson did not find a correlation between high stakes testing and a reduction in teacher cooperation. This was the only study we found that looked at a representative sample of teachers and their attitudes toward testing.

High stakes testing appears to have both positive and negative effects on students and teachers. Testing causes stress and frustration for both teachers and, reportedly, for students as well. Teachers reported decreased autonomy and ability to rely on their professional judgment. Interestingly, an increased sense of clear mission is the one positive attitudinal change to be documented from high stakes testing. Absent from this evidence is any direct study of how students have been affected both emotionally and academically by testing programs.

### What are the Emotional/Attitudinal Effects on Students with Disabilities?

Especially when it comes to graduation exit exams, one can assume that the effects on students with disabilities are at least as serious, if not more so, than on students without disabilities. McKinney (1983) discussed the possible repercussions for students with disabilities who, according to his study, have a limited probability of passing the test, even with modifications. The professional opinions of the teachers involved in the study were that students, especially those with mild mental disabilities, would experience frustration.

Many states leave the decision about the participation of students in high stakes tests up to the Individualized Education Program (IEP) team (see Thurlow et al., 1995). These teams need to take into account the possible negative effects of both participation and nonparticipation in the

testing programs. Again, little evidence exists about the real effects that high stakes testing has had on students with disabilities.

### Directions

As in other areas, the research into attitudes and school climate should focus more directly on students, particularly students with disabilities. Similarly, given the perceptions of educators, it might be beneficial to focus training efforts toward addressing the perceptions of negative effects and how to avoid these.

To date, there has been no research into the effects of high stakes testing on the relationship between special and regular education. If stakes for administrators and educators are high (for example, linked to promotions, bonuses, sanctions for poor performance, etc.) will regular educators see special education students as a threat, bringing down test scores and possibly resulting in the loss of funds or even jobs? Does high stakes testing affect the inclusion of students with disabilities in the regular classroom? Do special educators see high stakes testing as a threat to their autonomy and the ability to individualize the educational programs of students with disabilities? Do students with disabilities view high stakes testing as yet another barrier to their inclusion as members of our society, or do they see "just another test"?

## Costs Versus Benefits

So far in this report, we have examined the possible benefit or harm that may occur as a result of high stakes testing. In the next section, we will look at the costs versus the benefits of high stakes testing and the extent to which costs have been accurately estimated and taken into account during the development of high stakes testing programs.

### What are the Costs vs. the Benefits of High-Stakes Testing?

The issue of costs versus benefits of high stakes testing is not an area that has been thoroughly researched. Some advocates claim that tests can cause substantial improvement in schools at very little cost, with very little effort on the part of legislators, government, or the public. The idea seems to be that students and teachers will work harder to gain the rewards of scoring well on the tests, and avoid the consequences of scoring poorly.

The February 16th Daily Report Card (DRC), an on-line newsletter put out by the National Education Goals Panel, reported on a controversy about testing taking place in Virginia. According to the DRC, the governor was proposing a new testing program that would be used to determine district funding and could also be used to decide whether teachers and administrators should keep their jobs. Teachers complained that they were being asked to be

more accountable and to achieve higher standards without the tools and training to do so. An opponent of the proposal was quoted, saying, "Testing is education reform on the cheap. That's the problem. It's oversimplified." This section addresses whether testing reforms, particularly high stakes testing, are as cost-effective as proponents such as Popham (1987) claim.

According to Popham (1987), "If properly conceived and implemented, measurement-driven instruction currently constitutes the most cost-effective way of improving the quality of public education in the United States" (p. 679). A properly conceived and constructed test should be criterion-referenced, with defensible content containing a manageable number of targets designed for instructional illumination. He also provided for ample instructional support so that educators can make the best use of the tests, which he sees as "vehicles of instructional clarification" (p. 681). It is the task of the persons who design the tests to ensure that the content will drive both instruction and the curriculum. These things can all be accomplished without other elements of reform, such as better curriculum materials, or better paid and trained staff:

> if we were able to replace mediocre instructional materials with more potent, empirically proven alternatives, then pupils would surely benefit. Similarly, if we were to infuse into our current teaching force a host of well-paid, highly skilled teachers, we could surely expect major educational dividends. But such strategies, though they are surely *effective*, are very costly (p. 679).

Can a test be both "properly conceived and constructed" and still be less costly than other education reforms? Anderson (1977) brought up the subject of hidden costs of high stakes testing in his background paper prepared for the Minimal Competency Workshops sponsored by the Education Commission of the States and the National Institute of Education. Many of the hidden costs that Anderson outlined have yet to be taken into account when evaluating the testing programs of today. These include such things as test development, test administration, development and maintenance of regulatory mechanisms (bureaucracies), and compensatory programs to bring students who do not pass the minimum competencies up to the current standards. Anderson saw this last area, remediation, as containing the greatest hidden cost, and the evidence supports his prediction.

Potter and Wall (1992), in conducting a study of South Carolina's minimum competency testing program, found modest gains in student performance on minimum competency tests. Besides evidence mentioned before, that these gains may be attributable to factors other than actual learning (such as keeping children back a grade in order to avoid testing them for another year), the remediation efforts to raise the test scores had cost the state over $500 million. After spending that much money, the state still had no real evidence that the remediation had done much good.

Foshee, Davis, and Stone (1991) found that many students who in the past had failed a minimum competency test had subsequently passed without remediation. The study calls into question the cost-effectiveness of providing remediation to students who do not do well on a minimal competency exam.

Singer and Balow (1987), in evaluating California's proficiency law since 1980, found that students did better in regular English class than in remedial classes when retaking a minimum competency test. They were concerned by a shift in resources directed at students who were not doing well on the test, reducing the availability of advanced courses.

The evidence suggests that remediation programs not only are expensive, but often ineffective as well. The actual costs of state-administered high stakes tests are not known, especially as compared to the costs of other educational reforms.

### Who Benefits the Most from High Stakes Testing?

Anderson (1977) predicted that remediation programs would be too expensive, and suggested rewarding schools that perform well on the high stakes tests. Compensatory education programs, he believed, were emerging in the wrong direction, providing substantial incentive for schools to do poorly in order to get increased funds for remediation. Rewarding schools for good performance, however, has the added problem of diverting funds to schools that are already financially better off and doing well rather than giving the funds to schools that have greater needs. Anderson (1977) recommended that states work to reduce the financial inequities that exist between school districts as part of the overall effort to impose high stakes testing.

Socioeconomic status can have an impact on who benefits from high stakes testing. Tuma and Gifford (1995) found that higher graduation standards (not necessarily accompanied by high stakes testing) had affected the number of academic courses that college-bound students were taking, but had not affected students who were not college-bound. Potter and Wall (1992) found that overage students who had been withheld a grade prior to years where testing was to occur were more likely to be male and nonwhite. Reardon (1996) found that high stakes tests caused increased dropout rates only in schools with lower socioeconomic status. Herman and Golan (undated) found that:

> Correlations show that socioeconomic status is significantly and negatively related to the following: school attention to test scores, teachers' attention to testing and planning their instruction, and overall time devoted to test preparation activities . . . testing is more influential and exerts stronger effects on teaching in schools serving more disadvantaged students." (p. 57-58)

This study was not able to determine whether the increased effects on schools with lower SES resulted in better performance on the tests, so it is difficult to interpret these consequences as positive or negative.

Not all studies found a relationship between SES and performance on high stakes tests. Corbett and Wilson (1991) found that SES played a "surprisingly weak" role in explaining differences between districts.

It is very difficult to determine from the available evidence whether high stakes testing has different effects on students with low or with high SES. The evidence does suggest, however, that we need to take a closer look at this question to make sure that we are not using high stakes testing as a means to further inequities that already exist in our educational system.

### What About Portfolio and Authentic Assessments?

Portfolio and authentic assessments have been gaining popularity (Bond et al., 1996). Portfolio assessments are in-depth looks into a students' learning histories. They might include all of the assessments that the students take, as well as examples from their classroom work and other evidence of learning. Authentic assessments are designed to gain an in depth look at the students' performance level using tasks that are instructionally relevant to the child, and based on tasks that would normally be expected as part of a curriculum. Only recently have these types of assessments been used by states for high stakes purposes. These tests have many advantages over the usual multiple choice exam. They give more information about students, and are potentially more useful to teachers, and they measure higher order skills that are more difficult to assess with traditional paper and pencil tests. Portfolio and authentic assessments have several imposing disadvantages, however. In particular, they take more time to develop and implement. In addition, someone has to judge the students' responses and determine whether they meet the educational standards. The reliability of such judgments on a large-scale assessment program has yet to be established (Shepard, 1992).

Vermont was one of the first states to use portfolio assessments on a large-scale basis. According to Koretz, McCaffrey, Klein, Bell, and Stecher (1993), the intent of the assessment was to encourage high standards and good educational practices while maintaining local autonomy. Koretz et al. (1993) evaluated the 1992 Vermont Portfolio Assessment program and found disappointing reliability coefficients ranging from .33 to .43 (see Table 3).

| Table 3. Reliability Coefficients for the 1992 Vermont Portfolio Assessment | | |
|---|---|---|
| | Grade 4 | Grade 8 |
| Mathematics Best Pieces | .33 | .33 |
| Writing Best Piece | .35 | .42 |
| Writing Remainder | .34 | .43 |

From Koretz et al. (1993)

According to Koretz et al. (1993):

> The Vermont portfolio program faces substantial hurdles because of the unreliability of scoring documented here. Rater reliability is low enough to undermine the utility of 1992 scores for comparing groups of students (schools, districts, or other groups). . Even when scores are aggregated enough to produce estimates with small measurement and sampling error–for example, statewide reporting of average scores–low reliability threatens the usefulness for gauging trends in performance over time, because it remains uncertain how an increase in the reliability of ratings will affect the distribution of scores even if true performance remains constant. (p. 18)

Teachers, however, reported that they liked the portfolio assessments and thought that they were a valuable tool in gauging student progress, and many schools had expanded the portfolio program beyond the grades required by the state (Koretz et al. 1993, p. 1-2).

An even more ambitious testing effort to assess student learning for purposes of school accountability took place in England. Torrance (1993) looked at the effects of an authentic assessment program in Great Britain. The assessments used in the U.K. are designed to accompany a National Curriculum. The goal is to assess all children in the National Curriculum subjects at the ages of 7, 11, 14 and 16 through teacher assessment of course work (TA) and "standard assessment tasks" (SATs). Results are assembled into individual scores with 10 levels of achievement, with expected progress at one level per two years of schooling. Scores are reported to parents by subject and to the public by school. At the time of this study, implementation of this plan had begun only with the younger children (seven-year-olds).

> The aspiration was to produce tasks that would contribute to a system of assessment that was both formative and evaluative–aiding the learning of pupils by providing detailed information to teachers, and . . . providing directly comparable data on pupil and school achievement. At the same time, the tasks were meant to be as valid and 'user friendly' as possible . . . while also guiding the implementation of the National Curriculum through exemplifying what it should look like in action (examples of tasks include rolling toy cars down slopes of different gradients to investigate how far they travel and why; using

> dice in the context of a simple game to test computational skills; drawing and labeling a poster to illustrate how and why things grow). Thus, the 1990 pilot . . . was launched with the intention of piloting tasks that mirrored good primary education practice while at the same time yielding formatively useful information and comparable summative results. (p. 83)

The evaluation of this pilot project attempted to get responses from teachers in an informal, open-ended survey. Very few of the participating teachers responded. Of those who did respond, most said that they "were so exhausted and disillusioned with the whole business that they very nearly did not get in touch--they could hardly face spending yet more time on an experience they would rather forget" (p. 84).

The major complaint that teachers made in their responses was about their increased workload. According to Torrance, "the most commonly reported figure was two to three hours of extra work every evening for marking, record keeping, gathering resources, and planning the next day's work, plus six hours each weekend. All this was in addition to a full day's work in school beginning well before and ending well after the children's school day" (p. 85). This drain on teacher time had indirect effects on the rest of the school. Teachers of the seven-year-olds did not have time for assemblies, playground duties and other activities, leaving the burden of these activities to the other teachers in the school. Teachers reported that relationships with parents and students also were affected since teachers no longer had the time to welcome parents into the classroom, or offer extra help to students outside of class time. Other classroom duties were neglected as well, such as changing reading books and maintaining regular classroom routines. While students were engaged in the individual and small group tasks, the students who were not being assessed showed signs of boredom and stress from being ignored and subjected to "busy work." Torrance asserted, "It is clear, then, that teachers did not plan their curriculum, initiate a range of activities, and go on to assess in an opportunistic and "naturalistic" fashion, as some of the claims for classroom authenticity might lead us to believe. Rather, teachers treated assessment as a special activity, set apart from teaching, and they felt obliged to do this by the instructions they received" (p. 85).

Two major complaints about the implementation of the assessment program emerged from the survey. One was that the assessments contained so much new material that teachers felt "deskilled and overly dependent." The second was that the materials were too specific to be used in a naturalistic way, yet the specificity was deemed necessary in order to use the data for comparative purposes. Teachers reported that the materials were not flexible enough to expand upon when students showed genuine interest, and many students were anxious about the results rather than interested in the learning.

The assessment was "trimmed back" in 1991 and 1992, establishing what the author saw as a movement back toward paper-and-pencil tests. It is possible that we will see the same trend in the United States if testing programs have difficulty establishing reliable assessments.

One can easily criticize Torrance–certainly the self-selected nature of the survey could have been very skewed; however, on April 17, 1996, Education Week reported that the national assessments resulted in a major labor dispute in which teachers refused to administer the tests, which were seen as cumbersome and unwieldy (Bradley, 1996).

### What are the Costs Versus the Benefits of High-Stakes Testing for Students with Disabilities?

Researchers need to look not only at the inequalities that exist for students of different SES, but they also need to look at those that exist for students with and without disabilities. As mentioned before, students with disabilities stand a great chance of failing minimum competency tests, even when they are given accommodations such as increased time to take the tests, interpreters, or large print versions (McKinney, 1983; Safer, 1980). Bergquist et al. (undated) mentioned the need to provide increased supports for students with disabilities not only to do well on the tests, but also to maintain an appropriate program incorporating their need for vocational as well as academic subjects.

### Directions

Calculating the true costs of implementing statewide high stakes tests can be a daunting task. It would be useful for evaluators and researchers to have a framework to estimate the costs of testing programs, and to then balance these against the expected and realized benefits of the programs.

We also need to think more about inclusion of students with disabilities when developing any testing program. Portfolio assessments may make it easier for students with disabilities to participate in statewide testing programs, but it is still not clear whether this type of assessment can have the reliability needed to be used for statewide comparisons, graduation requirements, or other high stakes purposes, and the costs of such programs may be prohibitive.

## Recommendations and Conclusions

Special educators need to become more involved in development, implementation and evaluation of high stakes testing programs. The inclusion or exclusion of students with disabilities is a problem salient to both regular and special education researchers. The

following recommendations are just some of the many research and evaluation questions that should be taken into account in future investigations:

- Focus on the effects of high stakes testing on students rather than on schools and systems.
- Assess the effects of high stakes testing on the curriculum for both special and regular education.
- Assess the effects of high stakes testing on the dropout rates for both special and regular education.
- Study the effects of high stakes testing programs on students who are excluded from testing.
- Develop assessments that are more inclusive of students with disabilities (for example, including students with disabilities in state norming samples, and norming tests with some students using common accommodations such as extended time and Braille).
- Study the effects of high stakes testing on the relationship between regular and special education.
- Develop a framework for evaluating the costs versus the benefits of high stakes testing programs, particularly for alternative and authentic assessments.

The research on high stakes testing is inconclusive and results vary with the type of research questions asked, and the types of tests examined. The evidence suggests that teachers change the curriculum based on the tests, concentrating time and effort teaching to test contents and format. For students with disabilities, this may mean that less time is devoted to their vocational and other nonacademic needs, even though they are less likely to pass the test, even when given accommodations. The effects on student learning are largely unknown, but the evidence does suggest that increasing test scores themselves do not serve as evidence that students are learning more. Test scores can become inflated through teaching to the test, excluding or excusing students who may not perform well, and through increased dropout rates. High stakes testing seems to have a negative effect on the attitudes and workloads of teachers, but little is known about the effects on students themselves. States still do not take into account the full costs of high stakes testing programs, and claims that testing alone can cause major educational improvements have not been proven. Authentic and portfolio assessments hold promise, but it has not been established that these can be done well, efficiently and with sufficient reliability to be used as a large-scale comparative assessment method or for high stakes purposes like graduation.

The effects of high stakes testing on students with disabilities are harder to determine. Students with disabilities have not performed well on minimum competency tests, and it is unclear what effects other types of tests have had on their educational outcomes, choices and

futures. The needs of students with disabilities have been largely overlooked in the development and evaluation of statewide test-based reforms.

We do not advocate, as Corbett and Wilson (1991) suggest, that we abandon the practice of testing for educational outcomes altogether. It is our belief that schools need to be responsible to parents, taxpayers, and the community at large, for student outcomes. The evidence we have reviewed in this paper supports this position by demonstrating that exclusion of some students leads to abuses and statistical problems that make interpretation of test scores dubious at best. The evidence also tells us that testing alone is not a sufficient mechanism for effective school reform. The most properly conceived and implemented tests can only be effective if they support clear standards, solid curricula, committed and well-trained teachers, in schools with the resources and supports necessary to provide a world-class education to all students. Without these things, scores may continue to rise, but there is no reason to believe that our children will be any better educated or prepared for life than they were before.

## References


Allington, R., & McGill-Franzen, A. (1992a). Improving school effectiveness or inflating high-stakes test performance? ERS Spectrum, 10(2), 3-12.

Allington, R., & McGill-Franzen, A. (1992b). Unintended effects of reform in New York. Educational Policy, 6 (4), 397-414.

Anderson, B.D. (1977, September). The costs of legislated minimal competency requirements. A background paper prepared for the Minimal Competency Workshops sponsored by the Education Commission of the States and the National Institute of Education. (ERIC Document Reproduction Services No. ED 157 947)

Berger, N. & Elson, H.H. (1996, April). What happens when MCT's are used as an accountability device: Effects on teacher autonomy, cooperation and school mission. Paper presented at the annual meeting of the American Educational Research Association, New York.

Bergquist, C.C., Elzie, B., & Groves, L. (Undated). Evaluation of the impact and effectiveness of recent changes in Florida's graduation and competency test standards on the educational opportunities provided handicapped students. Paper presented at the First Annual Regional Conference of the Southeast Evaluation Association, Tallahassee, FL.

Bond, L.A., Roeber, E., & Braskamp, D. (1996). Trends in statewide student assessment. Washington, DC: Council of Chief State School Officers and NCREL.

Bradley, A. (1996, April 17). English reforms may offer model, report says. Education Week, p.5.

Catterall, J.S. (1987). Standards and school dropouts: A national study of the minimum competency test (CSE Technical Report 278). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Chin-Chance, S.A., Gronna, S.S., & Jenkins, A.A. (1996, March). Assessing special education students in a norm referenced statewide testing program: Hawaii State Department of Education. Paper presented at the meeting of the State Collaborative on Assessment and Student Standards (SCASS) Assessing Special Education Students, Washington, DC, sponsored by the Council of Chief State School Officers (CCSSO) March, 1996.

Corbett, H.D., & Wilson, B. (1990). Unintended and unwelcome: The local impact of state testing. Paper presented at the Annual Meeting of the American Educational Research Association, Boston.

Corbett, H.D., & Wilson B.L. (1991) Testing, reform, and rebellion. Norwood, NJ: Ablex.

Foshee, D.P., Davis, M.A., & Stone, M.A. (1991). Evaluating the impact of criterion-referenced measurement on remediation decisions. Remedial and Special Education (RASE), 12(2) 48-52.

Griffin, B.W., & Heidorn. (1996). An examination of the relationship between minimum competency test performance and dropping out of high school. Educational Evaluation and Policy Analysis, 18(3), 243-252.

Grossman, P.L., Kirst, M.W., & Schmidt-Posner, J. (1986). On the trail of the omnibeast: Evaluating omnibus education reforms in the 1980's. Educational Evaluation and Policy Analysis, 8(3), 233-266.

Hendrie, C. (1996, October 9). 109 Chicago schools put on academic probation. Education Week.

Herman, J.L., & Golan, S. (Undated). Effects of standardized testing on teachers and learning–another look (CSE Technical Report 334). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California.

Hodgkinson, H.L. (1992, June). A demographic look at tomorrow. Washington, DC: Institute for Educational Leadership, Center for Demographic Policy.

Hodgkinson, H.L. & Outtz, J.H. (1992, December). The nation and the states: A profile and data book of America's diversity. Washington, DC: Institute for Educational Leadership, Center for Demographic Policy.

Koretz, D.M., Linn, R.L., Dunbar, S.B., & Shepard, L.A. (1991). The effects of high-stakes testing on achievement: Preliminary findings about generalization across tests. Presented in R. L. Linn (Chair), Effects of High-Stakes Educational Testing on Instruction and Achievement symposium presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, Chicago.

Koretz, D., McCaffrey, D., Klein, S., Bell, R., & Stecher, B. (1993). The reliability of scores from the 1992 Vermont portfolio assessment program (CSE Technical Report 355). Los Angeles, CA: National Center for Research on Evaluation, Standards, and Student Testing (CRESST), University of California.

Madaus, G. (1988). The influence of testing on the curriculum. In L. Tanner (ed.), Critical issues in curriculum: 87th Yearbook of the NSSE Part 1. Chicago, IL: University of Chicago Press (ERIC Document Reproduction Service No. 263 183).

Mangino, E., Battaile, R., & Washington, W. (1986). Minimum competency for graduation (Publication Number 84.59). Austin, Texas: Austin Independent School District. (ERIC Document Reproduction Service No. 263 183)

Manzo, K.K. (1996, October 30). Phila. plan links student achievement, teacher pay. Education Week.

McKinney, J.D. (1983). Performance of handicapped students on the North Carolina minimum competency test. Exceptional Children,49, 547-550.

Morris, D.R. (1991, April). Structural patterns and change in grade retention rates: An aggregate analysis of data from a large urban school district, 1982-1989. Paper presented at the American Educational Research Association Conference, Chicago.

Mullis, I.V.S., Dossey, J.A., Campbell, J.R., Gentile, C.A., O'Sullivan, C., & Latham, A. (1994). NAEP 1992 trends in academic progress. Washington, DC: National Center for Education Statistics (NCES).

Office of Technology Assessment (OTA). (1992). Testing in American schools: Asking the right questions (summary report). Washington, DC.

Olson, J.F. & Goldstein, A.A. (1996). Increasing the inclusion of students with disabilities and limited English proficient students in NAEP. "Focus on NAEP' prepublication copy. Washington, DC: National Center for Education Statistics (NCES).

Popham, W.J. (1987). The merits of measurement-driven instruction. Phi Delta Kappan, 68 (9), 679-682.

Potter, D.C., & Wall, M.E. (1992, April). Higher standards for grade promotion and graduation: Unintended effects of reform. Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Reardon, S. (1996, April). Eighth grade minimum competency testing and early high school dropout patterns. Paper presented at the annual meeting of the American Educational Research Association, New York.

Rodgers, N., Paredes, V., & Mangino, E. (1991, April). High stakes minimum skills tests: Is their use increasing achievement? (ORE Publication Number 90.25). Paper presented at the annual meeting of the American Education Research Association, Chicago. (ERIC Document Reproduction Service No. ED 336 422).

Rottenberg, C., & Smith, M.L. (1990, April). Unintended effects of external testing in elementary schools. Paper presented at the annual meeting of the American Educational Research Association, Boston.

Safer, N.D. (1980). Implications of minimum competency standards and testing for handicapped students. Exceptional Children, 46, 288-290.

Salganik, L.H. (1985, May). Why testing reforms are so popular and how they are changing education. Phi Delta Kappan, 66(9), 607-610.

Shepard, L.A. (1990). "Inflated test score gains": Is it old norms or teaching the test? (CSE Technical Report 307). Los Angeles, CA: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L.A. (1992). Will national tests improve student learning? (CSE Technical Report 342). Los Angeles: University of California, Center for Research on Evaluation, Standards, and Student Testing.

Shepard, L.A., & Dougherty, K. (1991, April). Effects of high-stakes testing on instruction. Paper presented at the annual meeting of the American Education Research Association and the National Council on Measurement in Education, Chicago.

Sinclair, M., Christenson, S., Thurlow, M. & Evelo, D. (1994). Are we pushing students in special education to drop out of school? Research Policy Brief, 6 (1). Minneapolis, MN: University of Minnesota, Research and Training Center on Residential Services and Community Living.

Singer, H., & Balow, I.H. (1987). Proficiency assessment and its consequences. Riverside, CA: University of California. (ERIC Document Reproduction Service No. ED 290 127)

Thurlow, M.L., Ysseldyke, J.E., & Anderson, C.L. (1995, May). High school graduation requirements: What's happening for students with disabilities? (Synthesis Report 20). Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes.

Torrance, H. (1993). Combining measurement-driven instruction with authentic assessment: Some initial observations of national assessment in England and Wales. Educational Evaluation and Policy Analysis, 15(1), 81-90.

Tuma, J., & Gifford, A. (1995). Higher graduation standards and their effect on the course-taking patterns of college- and non-college-bound high school graduates, 1969-1987. Berkeley, CA: MPR Associates (ERIC Document Reproduction Service No. ED 318 767).

Vitello, Camilli, & Molenaar. (1987). Performance of special education students on a minimum competency test. Diagnostique, 13 (1), 28-35.

Walstad, W.B. (1984, May/June). Analyzing minimal competency test performance. Journal of Educational Research, 77(51) 261-266.

Zlatos, B. (1994). Don't test, don't tell: Is 'academic red-shirting' skewing the way we rank our schools? The American School Board Journal, 181(11) 24-28.

# Appendix

## Summary of Literature Reviewed for This Report

## Effects on Curriculum and Instruction

| Study | Type of Test | Method | Subjects | Results |
|---|---|---|---|---|
| Berger & Elson (1996). What happens when MCTs are used as an accountability device: Effects on teaching autonomy, cooperation and school mission (also in Attitudes and School Climate). | Graduation exit exam | Survey. Compared responses of teachers in high stakes programs (graduation exit exam) to teachers in low stakes programs. | National Schools and Staffing Survey (SASS), Dept. of Ed. 1987. Representative national sample of teachers from low and high-stakes states. | Loss of teacher autonomy; increased clarity of mission; no effect on teacher cooperation; increased emphasis on basics. |
| Bergquist, Elzie & Groves (undated). Evaluation of the impact and effectiveness of recent changes in Florida's graduation and competency test standards on the educational opportunities provided handicapped students. | Increased standards and graduation requirements. Did not report on the effects of the graduation exit exam. | Program evaluation. Student records, surveys, and qualitative observations. | 93 onsite visits statewide. 300 interviews from all levels of system. Reviewed more than 100 student records. Surveyed 500 knowledgeable people. Onsite visits to nine representative districts. | Students with disabilities had difficulty meeting the higher standards and incorporating nonacademic subjects into the curriculum. |
| Grossman, Kirst, & Schmidt-Posner (1986). On the trail of the omnibeast: Evaluating omnibus education reforms in the 1980s. | California education reforms including increased graduation requirements, increased college entrance requirements. | Examined course offerings from 1982 to 1985. | Sampled course offerings in 200 school districts statewide. | Increased offerings in academic areas, especially math, science, and advanced placement; decreased offerings in industrial arts, home economics and business ed. |
| Herman & Golan (undated). Effects of standardized testing on teachers and learning--another look. | Standardized testing in nine states. | Survey | Teachers chosen in matched pairs from demographically similar schools that had shown substantial improvement on standardized tests and those which had not. Total of 24 pairs. | Substantial time and effort devoted to testing; pressure from schools to improve scores; modification of curriculum; greater impact in lowest SES schools. Gains not clear as to real improvement or teaching to the test. |
| Rottenberg & Smith (1990). Unintended effects of external testing in elementary schools (also in Attitudes and School Climate). | ITBS used for evaluation of principals and making curriculum decisions. Media coverage of test scores. | Interviews with teachers. Dual case study design. | 19 teachers in two school districts. Observation of four local teachers. | Negative effects on the attitudes of both teachers and students. Reduced time for ordinary instruction. |
| Shepard (1990). "Inflated test score gains." Is it old norms or teaching to the test? | All | Survey of state directors of testing regarding narrowing of the curriculum and teaching to the test. | State testing directors in the 46 states that had testing programs. | Found instances of teaching to the test. Could not quantify how teaching to the test had affected test scores. Recommended new tests every year. |
| Shepard & Dougherty (1991). Effects of high stakes testing on instruction. | Standardized tests (two states, unspecified). Third, fifth, and sixth grade testing programs. | Survey | Teachers in districts with high stakes testing. 360 teachers responded (42% return rate). | Greater emphasis on basic skills; neglect of nontested material; inordinate time on test preparation throughout year; clearer instruction goals. |





BEST COPY AVAILABLE

## Effects on Student Learning

| Study | Type of Test | Method | Subjects | Results |
|---|---|---|---|---|
| Allington & McGill-Franzen (1992a). Improving school effectiveness or inflating high-stakes test performances? | Third grade testing in reading and mathematics. | Examination of student records, using students' scores by kindergarten cohort rather than grade. | Seven schools (low, moderate and high poverty levels). 11-70 subjects in each school. Total subjects not reported. | Data suggest that reported test score gains result from retention and referral to special education. |
| Allington & McGill-Franzen (1992b). Unintended effects of educational reform in New York. | Third grade testing in reading and mathematics. | Analyses of trends in special education referral and grade retention. | Seven schools (low, moderate and high poverty levels). 11-70 subjects in each school. Total subjects not reported. | Data suggest increasing trends in special education referral and grade retention during increased stakes. |
| Catterall (1987). Standards and school dropouts: A national study of the minimum competency test. | Graduation exit exam. | Interviews with educators and students in four states with graduation exit exams. | 736 students. Six state test directors. 13 district test coordinators. 18 school principals. 21 school counselors. | Professionals did not believe that the exit tests had affected dropout rates. Students, while supportive of testing in general reported more effects on dropout rates. Found association between competency test failure and reduced beliefs that the student would finish school. |
| Corbett & Wilson (1991). Testing, reform and rebellion (also in Attitudes and School Climate). Corbett & Wilson (1990). Unintended and unwelcome: The local impact of state testing. | Graduation exit exam (Maryland). Exam to identify students in need of remediation (Pennsylvania). | Compared high-stakes state (MD) to low stakes (PA). Qualitative interviews, extended site visits, and surveys. Twelve sites visited (six from each state). | Over 250 educators. | Greater impact in high stakes state; narrowed curriculum; greater emphasis on basic skills; neglect of nontested subjects; increased preparation time for tests; increased clarity of educational goals. |
| Koretz, Linn, Dunbar, & Shepard (1991). The effects of high-stakes testing on achievement. Preliminary findings about generalization across tests. | Third grade high-stakes. Specific test and location not revealed. | Compared scores on high-stakes test to other tests of achievement in the same academic areas. | Third grade student in large, high-poverty district. 620-753 students. | Little correlation between the high stakes test and other tests of achievement, suggesting that reported improvements in test scores were due to learning specific to test rather than general academic skills. |
| Mangino, Battaile, & Washington (1986). Minimum competency for graduation. | Graduation exit exam. | Analyzed data from the Texas graduation exit exam 1984-1985. | 184 students in reading and 115 in math (total number of students receiving waivers). | Problem areas found, including taking the test many times (up to 4), waivers, special education exemptions. |
| Morris (1991). Structural pattern and change in grade retention rates: An aggregate analysis of data from a large urban school district, 1982-1989. | Test used initially to identify weak students, later used to identify weak programs (from low to high stakes). | Analyzed data from Florida's minimum competency testing programs. | All students taking the exam. | Increased retention rates due to school restructuring and increased standards. |

## Effects on Student Learning (cont.)

| Study | Type of Test | Method | Subjects | Results |
|---|---|---|---|---|
| Potter & Wall (1992). Higher standards for grade promotion and graduation: Unintended effects of reform. | Graduation exit exam, grade promotion. | Longitudinal study of effects of SC graduation, promotion exams, 1985-91. | All students taking SC exams. | Students were more likely to be retained (were overage for their grade) as a result of high-stakes testing. |
| Griffin & Heidorn (1996). An examination of the relationship between minimum competency test performance and dropping out of high school. | Graduation exit exam. | Examined the effects of graduation exit exam on dropout rates in Florida. | Cross-sections, random sample of students in high school from 14 school districts, grades 10,11, and 12. N=76,664 students in 75 high schools. | Students who did not pass the MCT test were not more likely to drop-out, regardless of SES, other factors. Exception: students with high GPAs who failed the exam were more likely to drop out. |
| Reardon (1996). Eighth grade minimum competency testing and early high school dropout patterns. | Promotion from eighth grade to ninth grade. | Examined data from National Educational Longitudinal Survey (NELS). | Focused on MCT in eighth grade. Nationwide representative sample. | Increased dropout rates in programs with eighth grade MCT. |
| Walstad (1984). Analyzing minimal competency test performance. | Test used to gauge district/student performance. | Used an educational production function model based on prior research and two years of MCT data. Surveys of district administrators. | District administrators. Used data from all students taking MCT in state. | Only pretesting (practicing the test) had a significant effect on student performance. Changes in curricula and teacher training had no effect on scores. |
| Chin-Chance, Gronna & Jenkins (1996). Assessing special education students in a norm referenced statewide testing program. | SAT8 (Stanford Achievement Test 8). Used for district comparisons. | Developed separate norms, percentile ranks for students with disabilities. | Students with disabilities (all taking exam). | Were able to develop separate percentile ranks. While students with disabilities scored lower than students without disabilities, difference scores showed students with disabilities improved as much or more than students without disabilities. |
| McKinney (1983). Performance of handicapped students on the North Carolina minimum competency test. | Graduation exit exam. | Analyzed data on how students with disabilities performed with and without accommodations (not under control of experimenter) | 3,043 students with disabilities. | Students with disabilities performed poorly on the tests, even when given accommodations. Persons with mild MR received the most benefit from accommodations but were nevertheless the least likely group to pass the test. |
| Safer (1980). Implications of minimum competency standards and testing for handicapped students. | Graduation exit exam. | Examined scores of students with disabilities taking graduation exit exam. | All students with disabilities taking the exam in 1977. | Students with disabilities were not likely to pass the exam. |
| Vitello, Camilli & Molenaar (1987). Performance of special education students on a minimal competency test. | Graduation exit exam. | Analyzed scores of students with disabilities taking the NJ proficiency exam. | All 4,299 students with disabilities who took exam 1986-87 (40% of the ninth grade handicapped population). | Students with disabilities were not likely to pass. |

## Effects on Attitudes and School Climate:

| Study | Type of Test | Method | Subjects | Results |
|---|---|---|---|---|
| Berger & Elson (1996). What happens when MCTs are used as an accountability device: Effects on teaching autonomy, cooperation and school mission (also in Curriculum). | Graduation exit exam. | Survey. Compared responses of teachers in high stakes programs (graduation exit exam) to teachers in low stakes programs. | National Schools and Staffing Survey (SASS), Dept. of Ed. 1987. Representative national sample of teachers from low and high-stakes states. | Loss of teacher autonomy; increased clarity of mission; no effect on teacher cooperation; increased emphasis on basics. |
| Corbett & Wilson (1991). Testing, reform and rebellion (also in Curriculum). | Graduation exit exam (Maryland). Exam to identify students in need of remediation (Pennsylvania). | Compared high-stakes state (MD) to low stakes (PA). Qualitative interviews, extended site visits, and surveys. Twelve sites visited (six from each state). | Over 250 educators. | Greater impact in high stakes state; narrowed curriculum; greater emphasis on basic skills; neglect of nontested subjects; increased preparation time for tests; increased clarity of educational goals. |
| Rottenberg & Smith (1990). Unintended effects of external testing in elementary schools (also in Student Learning). | ITBS used for evaluation of principals and making curriculum decisions. Media coverage of test scores. | Interviews with teachers. Dual case study design. | 19 teachers in two schools districts. Observation of four local teachers. | Negative effects on the attitudes of both teachers and students. Reduced time for ordinary instruction. |
| Rodgers, Paredes, & Mangino (1991). High Stakes minimum skills tests: Is their use increasing achievement? | Graduation exit exam. | Used multivariate analysis of variance (MANOVA) on the year of the test, GPA, year X GPA interaction, and GPA squared. Expected basic skills to increase while higher order skills decreased. | 12,404 11th grade students in the Austin Independent School District. Special education and LEP students were not included in the study. | Basic skills increased while higher order skills remained the same. |

## Costs/Benefits of High Stakes Testing

| Study | Type of Test | Method | Subjects | Results |
|---|---|---|---|---|
| Foshee, Davis, & Stone (1991). Evaluating the impact of criterion-referenced measurement on remediation decisions. | Preparation exam designed to identify students in need of remediation in the eighth grade in preparation for the graduation exit exam in the 10th grade. | Examined scores of students to determine which ones would meet the future criteria for remediation. | All students in county school system who had completed both testing session (preparation exam and exit exam). 1,310 students. | Many students who would have been identified for remediation passed the test without remediation. Calls to question the ability of the criterion referenced test to identify students in need of remediation. |
| Koretz, McCaffrey, Klein, Bell, & Stecher (1993). The reliability of scores from the 1992 Vermont Portfolio assessment program. | Portfolio assessment. State wanted to use data for comparative purposes. | Analysis of data for 1992-93 school year. | All students participating in the Vermont portfolio program. | Low reliability in mathematics and writing. |
| Singer & Balow (1987). Proficiency assessment and its consequences. | Graduation exit exam. | Survey, analysis of state data on testing. | Survey of students in California and New York taking graduation exit exam. | Students in regular English did as well as students in remedial English. Costs of locally developed programs in California were high and unreliable compared to New York which had a unified testing system. |
| Torrance (1993). Combining measurement-driven instruction with authentic assessment. Some initial observations of national assessment in England and Wales. | Authentic assessment-proficiency exam. Grade school. | Survey–very low response rate. | Teachers participating in the project. | Inordinate time on test preparation; neglect of other classroom duties. |

The College of Education
& Human Development

UNIVERSITY OF MINNESOTA